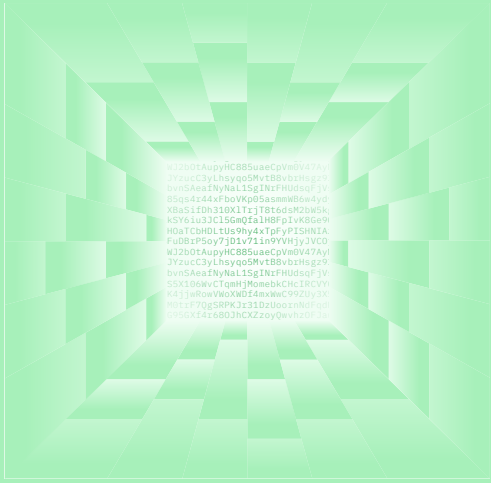


利用定制化生成式 AI 实现精准发力

不同于以往的任何技术，生成式 AI 正在迅速颠覆商业和社会形态，迫使企业领导者刻不容缓地重新思考其假设、计划和战略。

为了帮助 CEO 们掌握快速变化的形势，IBM 商业价值研究院 (IBM IBV) 发布了一系列有针对性、基于研究数据的生成式 AI 指南，涵盖数据安全、技术投资策略和客户体验等主题。

这是本指南的第十八部分，重点关注 AI 模型优化。

生成式 AI 的多样性

ChatGPT 让人们误以为自己都是 AI 专家，但这种表面的简单性掩盖了生成式 AI 领域的复杂性。CEO 在构建 AI 模型组合时必须考虑这些复杂因素。

生成式 AI 模型有多种类型，每种模型的功能、效果和成本都大相径庭。模型的所有权、开发方式和训练数据集的大小都是影响不同应用场景下模型选择的重要因素。

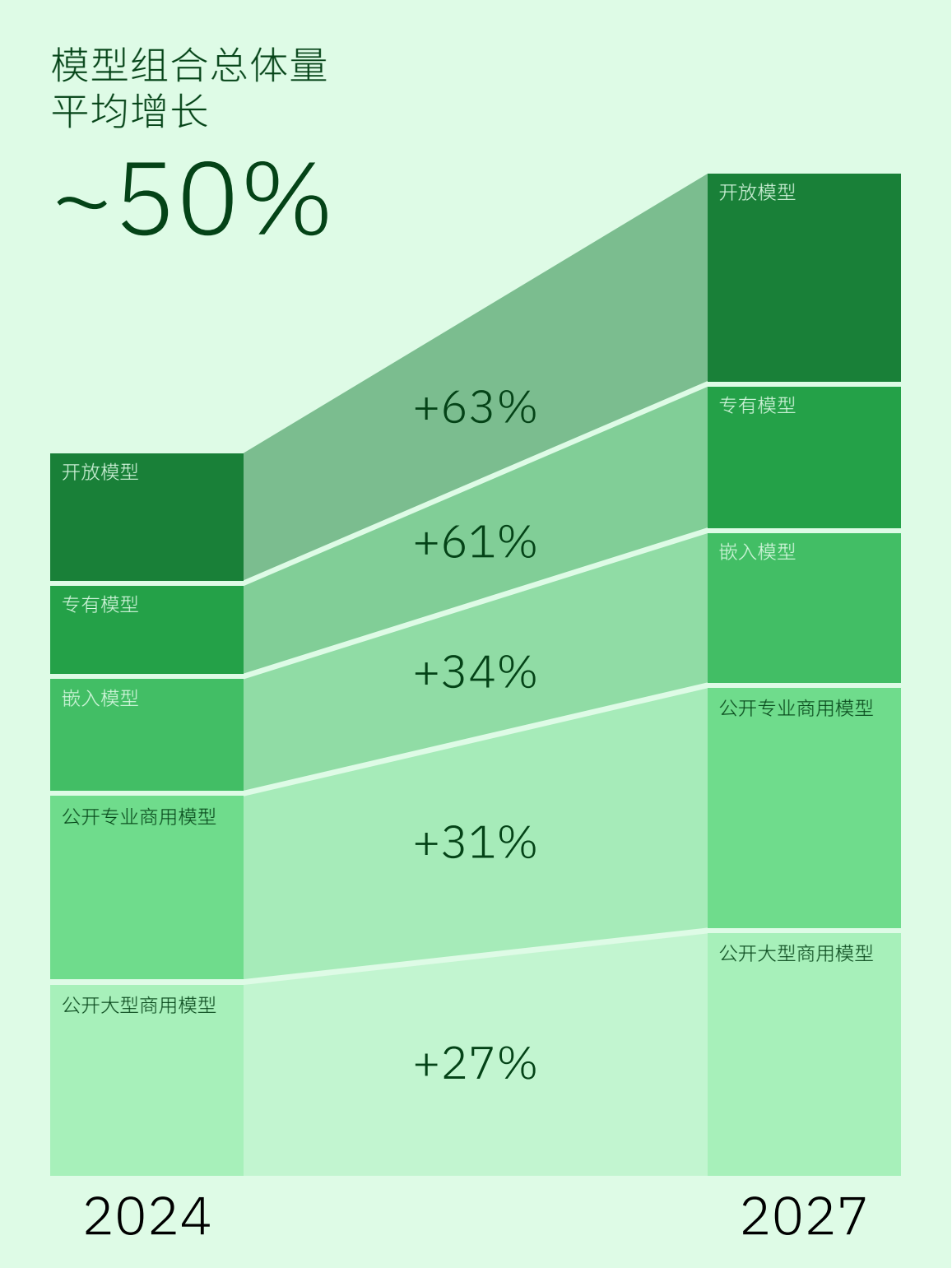
由于训练单个大语言模型 (LLM) 需要海量数据和资源，因此围绕生成式 AI 讨论的一个主要问题就是规模。因此，许多 CEO 都在考虑是否应为其业务大规模扩展大型 AI 模型，或者还是应当开发针对特定用途的小型专业 AI 模型。

答案是需要双管齐下。

许多组织已经开始这样做了。目前，一家典型的组织使用 11 种生成式 AI 模型，并预计会在未来三年内将其模型组合扩大约 50%。

为什么需要如此多的模型？因为每个应用场景都有各自的需求和限制，而不同的业务问题也需要不同类型的模型。

例如，图像编辑或数据分析等高度专业化的任务需要基于小型专业数据集进行训练的生成式 AI 模型。敏感或专有的工作则需要能够保证机密性的生成式 AI 模型。而对于文本生成等常规性任务，就可能需要在尽可能大的数据集上训练生成 AI 模型。



尽管团队应当深入理解不同 AI 模型的具体差异，但 CEO 也需要认识到为生成式 AI 的每种应用场景选择合适模型的重要性。理解哪些因素会影响成本、环境足迹和业务价值有助于优化 AI 模型组合的性能，并为团队提供超越竞争对手所需的利器。

IBM 商业价值研究院甄别出了每位领导者都需要了解的三个要点：

1. 不存在万能的 AI 模型。
2. 生成式 AI 成本完全可控。
3. 生成式 AI 的优势转瞬即逝。

现在，每位领导者都需要采取以下三项行动：

1. 为不同的团队提供不同的工具。
2. 找到生成式 AI 的最佳平衡点。
3. 让模型更高效地运作。

1. 敏捷性 + 生成式 AI

需要了解的事项

不存在万能的 AI 模型

生成式 AI 可以帮助组织更加精准、敏捷地加快行动——但前提是要在合适的环境中针对合适的目标运行合适的 AI 模型。

尽管技术高管最有能力决定在何处使用哪种生成式 AI 模型，但了解不同模型类型的优缺点以及竞争对手的走向，有助于 CEO 做出更明智的投资决策。

模型类型	示例和特点
开放模型	**Granite、Mistral** – 训练方式因规模和专业化程度而异 – 注重透明度和职责划分 – 不同公司/模型具有不同的开放程度 – 更具创新潜力
专有模型	**企业定制开发模型** – 训练成本由企业承担 – 更有效地控制范围和数据 – 更具差异化潜力
嵌入模型	**Models in SAP Joule、Salesforce Einstein 和 Adobe Firefly** – 集成到现有企业软件中 – 通常利用现有模型作为软件产品功能 – 通常不可独立使用
公开专业商用模型	**Google Med-PaLM** – 基于大型专有数据集进行训练 – 注重深度和专业化 – 通常不透明 – 具有一定的差异化潜力
公开大型商用模型	**GPT-4** – 基于海量数据集进行训练 – 注重广度和深度 – 通常不透明 – 差异化潜力有限

需要了解的事项

不存在万能的 AI 模型（续）

我们从调研中收集到了一些有用的数据。例如，在一家典型组织使用的模型组合中，公开可用的大型商用模型（如 GPT-4）仅占约四分之一。公开专业商用模型（如 Google Med-PaLM）占 23%，开放模型（如 Granite 和 Mistral）占 16%，嵌入模型（如 SAP Joule、Salesforce Einstein 和 Adobe Firefly）占 14%，企业定制开发的专有模型占 11%。除此之外，其他模型占 12%。

在为工作流程选择生成式 AI 模型时，模型规模是技术高管的首要考虑因素之一。大型模型基于数千亿个参数进行训练，可提供更广泛和深入的专业知识，处理更复杂的任务，但其价格更高，碳足迹也更大。相比之下，较小的专业模型通常基于数百亿个参数进行训练，可以更精准、快速、高效地处理特定任务，例如将代码或内容翻译为特定语言。

模型所有权是另一个重要考虑因素。尽管公开商用生成式 AI 模型非常受欢迎，约占企业 AI 模型组合的一半，但存在其局限性。任何企业都可以购买或获准使用这些模型，因此所有企业都在使用相同的数据，也就无法有效建立差异化优势。公开模型可以帮助团队更加快速高效地工作，但这些模型在公共云上运行，因此无法为企业提供处理关键任务所需的隐私和控制。

而这正是企业专有生成式 AI 模型的用武之地。此类模型是由使用模型的企业开发、拥有和控制的，因此企业可以决定用哪些数据来训练模型，从而减少模型及其输出受到错误信息污染的可能性。这些专有模型还为技术高管提供了更高的灵活性，可决定是在本地环境还是云端运行模型，以及如何存储或使用用户提供的信息来调优模型性能，从而减少私有或敏感数据被不当使用或分享的风险。这是一项至关重要的能力，因为误用、隐私和准确性是高管在选择生成式 AI 模型时最关心的问题。

开放生成式 AI 模型是在开源开发者社区的帮助下透明地构建的，规模可大可小，也可以解决这些问题。由于此类模型是公开构建的，因此用于训练模型的数据是公开透明的，并且经过严格审查，可迅速识别和应对各种风险及问题，例如输出是否侵犯知识产权或版权。随后，企业可以修改和定制这些基础模型，以加速创新、提升性能以及建立对生成式 AI 的信任。

嵌入生成式 AI 模型的来源多样，完全嵌入到 SAP、Adobe 和 Salesforce 等平台或软件中，可满足软件功能范围内的特定需求。此类模型可为所支持的产品提供增值，但无法单独使用。

在未来三年内，生成式 AI 模型的采用将大幅增长，其中最具增长潜力的是开放模型。平均而言，受访企业高管预计其 AI 模型组合中的开放模型将增长 63%，其背后的驱动力包括灵活性、透明性和定制化需求。同时，受访高管还预计更加可靠且易于扩展的大型商用模型的使用量将增长 27%，可处理更专业化任务的专业商用模型的使用量将增长 31%。同样在未来三年内，受访高管预计专有模型的使用量将增长 61%，嵌入模型的使用量将增长 34%。

1. 敏捷性 + 生成式 AI

需要采取的行动

为不同的团队提供不同的工具

评估基础模型组合，并确定与战略工作流的契合度。投资部署大型生成式 AI 模型以提高生产力，并利用专业模型来处理更有针对性的任务。

建立生成式 AI 全景图。了解不同类型的生成式 AI 模型的区别，包括大语言模型、企业定制开发的专有模型、开放模型等。做好充分准备，针对不同用途投资部署不同的模型。

绘制 AI 模型地图。要求 AI 高管创建全面的生成式 AI 模型目录，涵盖组织内使用的所有模型的用途、功能和性能指标，并确保该目录定期更新以反映 AI 领域的变化。

找到最佳匹配。确保团队根据其优势、劣势和特点来将生成式 AI 模型与合适的工作流相匹配。识别差距——但如果一本字典就能解决问题，就不要使用一整套百科全书。

2. 成本 + 生成式 AI

需要了解的事项

生成式 AI 成本完全可控

CEO 知道其组织需要生成式 AI——但成本多少？随着生成式 AI 逐渐渗透到企业的各个领域，企业高管表示首先考虑的是如何在规模化应用中实现成本效益，以便在不同场景中选择合适的模型。

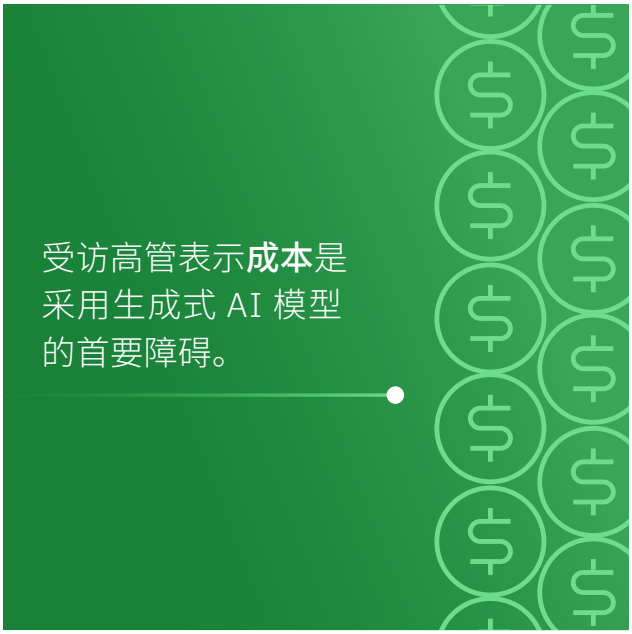
就面临的障碍而言，63% 的受访高管表示模型成本是最担忧的问题，而 58% 的受访高管则认为模型复杂性是最担忧的问题。

为什么成本如此重要？因为成本会因所使用的模型而存在很大的差异。例如，更大的模型需要更多的数据存储和计算资源，这可能导致更高的云计算费用。此外，大型模型还需要更频繁的更新、调优和维护，这也会增加人力成本。相比之下，专业模型则具有较低的计算、数据存储和能源成本，并且可减少组织 AI 模型组合对环境产生的影响。而且专业模型的部署速度更快，维护需求更少，因此可降低人力成本。

根据具体任务选择最合适的模型规模对于帮助组织管理生成式 AI 成本至关重要。例如，长篇写作、高风险决策和研究假设测试等复杂任务需要多种技能和高精度性，因此也就需要成本更高的大型模型。而更具成本效益的专业模型则更适合处理更具针对性的任务，尤其是速度和效率至关重要的任务，例如实时聊天支持、垃圾邮件检测、数据增强和原型设计。团队还可以利用链式推理等先进技术，将复杂工作分解为专业模型能够处理的小任务，从而减少对成本较高的大语言模型的依赖。

随着技术的成熟，专业模型将能够处理更广泛的任务，让组织有机会实现更精细的成本管理。借助“针对特定用途的专用”模型，即根据特定需求和目标进行设计、训练和验证的模型，团队可以仅为每项任务使用所需的资源。如果使用大型模型来训练更具针对性的专业模型，企业还可以提高模型开发的成本效益。

在不久的将来，企业高管或许可以通过企业生成式 AI 控制中心来改善成本管理，从而简化关于应为每项任务使用哪种模型的决策。通过添加一个用户友好的体验层，将整个 AI 模型组合中的模型、助手和提示连接起来，企业高管可以实施成本控制，同时确保安全、隐私和合规性，从而让每位员工每次都能高效地使用模型。



2. 成本 + 生成式 AI

需要采取的行动

找到生成式 AI 的最佳平衡点

发掘多样性的价值：为每项任务使用适当规模的生成式 AI 模型，有效控制成本并提高整体 AI 投资回报率。

培养与模型无关的思维方式。保持灵活性，采用针对价格和性能进行了优化的模型，在准确性、资源使用和速度之间取得平衡。

追求效率设计。根据部署环境来调整模型范围：针对移动和实时应用，优先选择速度更快的小型专业模型，而针对复杂的高精度任务，则优先选择大型模型。

削减不必要的开支。为每一个生成式 AI 部署建立明确的性能指标和对标。使用数据驱动的洞察，了解生成式 AI 可在哪些领域实现预期价值，以及需要在哪些方面控制成本。

3. 竞争力 + 生成式 AI

需要了解的事项

生成式 AI 的优势转瞬即逝

生成式 AI 当前带来的竞争优势，未来可能只是基本要求。随着团队获得更丰富的生成式 AI 经验以及模型变得更加智能，CEO 必须将持续改进列为首要任务。

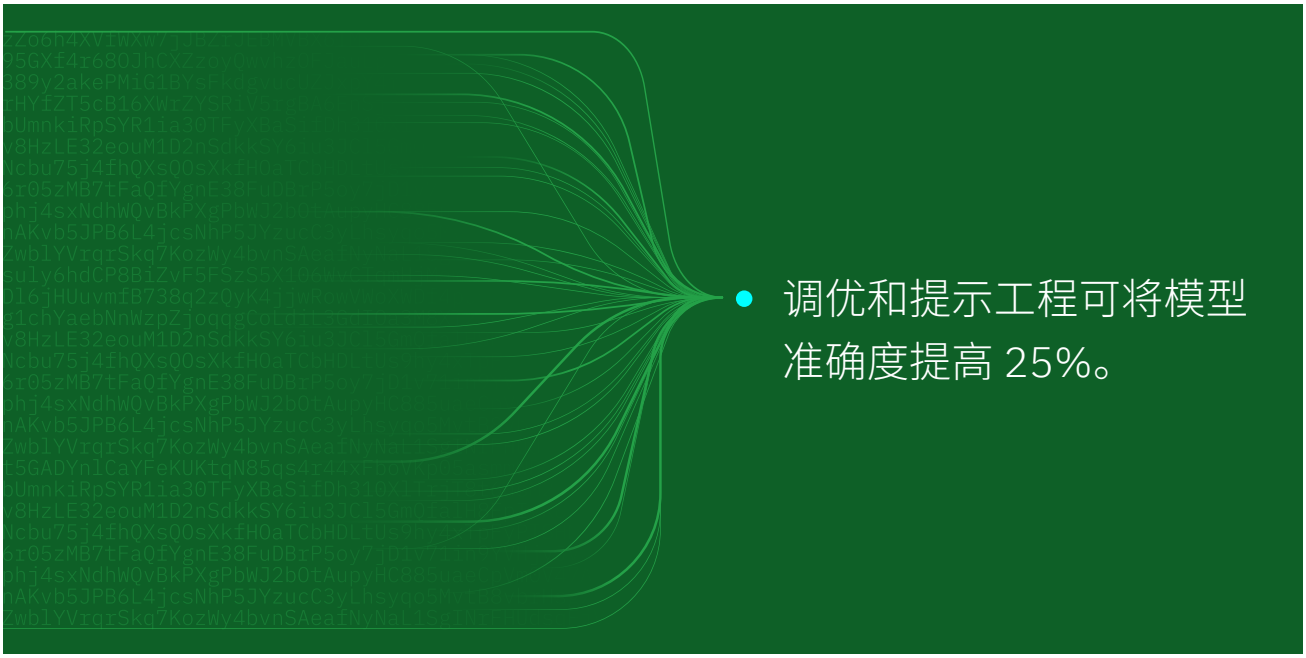
致力于持续优化的组织有望实现显著的绩效提升。根据 IBM 商业价值研究院的研究，对于使用调优或提示工程技术的组织，其模型输出的准确度要比其他组织高出约 25%。更高的准确度有助于改进预测能力、资源分配和个性化体验，从而最终转化为更高的盈利水平。

然而，只有 42% 的受访高管表示始终会使用提示工程技术（即通过设计输入来确保生成符合预期的输出）来提高模型准确度。

但模型优化只是解决方案的一部分。随着模型组合的不断发展，模型治理也必须同步演进。这就需要定期更新企业的内部管理方式，以便有效管理和控制模型库，以及明确哪些人员有权开发、训练和调优模型。组织还需要通过清晰的流程来跟踪模型绩效指标，处理模型漂移（即模型准确性随着时间的推移而下降），以及纠正模型输出中的偏差。最重要的是，团队还需应对快速变化的法规以保持合规性。

组织还需要持续改进其 AI 基础架构（即混合云战略），以便开发和采用更强大的 AI 模型。随着数据量和模型复杂性的增加，技术基础架构必须能够处理更高的负载。接下来是扩展的问题。随着越来越多的团队开始使用各种不同形式的生成式 AI，组织需要扩展其基础架构或云环境来满足日益增长的需求。

那么当前的现实情况是什么样的？目前，至少有一半的组织正在专注于优化网络基础架构、加速数据处理或部署分布式计算。总体而言，63% 的受访高管表示其组织正在使用至少一种基础架构优化技术。



3. 竞争力 + 生成式 AI

需要采取的行动

让模型更高效地运作

不要满足于早期的成功。持续推动团队利用最新的 AI 技术和基础架构来提升模型性能并超越竞争对手。

提高生成式 AI 的标准。将企业数据整合到私有云或本地部署环境中的现有生成式 AI 模型中，打造组织的独有优势。使用调优、提示工程和其他优化技术，始终保持领先竞争对手三步。

打造面向未来的 AI 基础架构。投资部署基于云的服务或专用硬件以及开放框架，以便持续利用 AI 驱动的创新来推动变革。

避免被边缘化。建立清晰的治理框架，加速推动生成式 AI 的应用。质疑自身在监管准备方面的假设，成为最严格的自我审查者。

IBM 商业价值研究院

CEO 生成式 AI 行动指南

AI 模型优化

本报告分析所依据的统计数据来自 IBM 商业价值研究院联合牛津经济研究院开展的一次专项调查。这项调查于 2024 年 6 月询问了 200 名美国高管对 AI 模型优化的看法。

IBM 商业价值研究院

IBM 商业价值研究院 (IBM IBV) 成立二十多年来，凭借IBM在商业、技术和社会交叉领域的独特地位，每年都会针对成千上万高管、消费者和专家展开调研、访谈和互动，将他们的观点综合成可信赖的、振奋人心和切实可行的洞察。

需要 IBV 最新研究成果，请在 ibm.com/ibv 上注册以接收 IBV 的电子邮件通讯。您可以在 Twitter 上关注 @IBMIBV，或通过 https://ibm.co/ibv-linkedin 在 LinkedIn 上联系我们。

访问 IBM 商业价值研究院中国官网，免费下载研究报告：https://www.ibm.com/ibv/cn

国际商业机器（中国）有限公司
北京市朝阳区金和东路 20 号院 3 号楼
正大中心南塔 12 层
邮编：100020

美国出品 | 2024 年 9 月

WPXGRV7D-ZHCN-01

扫码关注 IBM 商业价值研究院

官网　　微博　　微信公众号